Análise Comparativa de Modelos de Regressão para Previsão do IDHM de Cidades Brasileiras Usando Aprendizado de Máquina

Afonso Simão de Gois, Mikael Johnatan da Silva, Pedro Figueira e Amanda Gondim

Termos de índice— **Aprendizagem de Máquina**

***Resumo*— Este artigo apresenta uma análise comparativa entre diferentes algoritmos de regressão de aprendizado de máquina aplicados à previsão do Índice de Desenvolvimento Humano Municipal (IDHM) de cidades brasileiras. Foram examinadas técnicas simples, como Regressão Linear e K-Nearest Neighbors (KNN), e modelos mais sofisticados, como Árvores de Decisão, Florestas Aleatórias e Árvores Extremamente Aleatórias. Utilizou-se um conjunto de dados socioeconômicos que abrange perfis diversos de cidades, avaliando a capacidade de cada modelo em capturar as complexas relações entre as variáveis. Técnicas de pré-processamento, como normalização, remoção de outliers e redução de dimensionalidade via PCA, foram aplicadas para melhorar o desempenho dos modelos. Dentre os algoritmos testados, o modelo de Árvores Extremamente Aleatórias obteve o melhor resultado, com um R^2 de 0,8816, superando os demais. Os resultados indicam que modelos baseados em árvores, especialmente as Árvores Extremamente Aleatórias, são mais adequados para tarefas de regressão complexas, como a previsão do IDHM, proporcionando maior precisão e robustez quando aliados a um tratamento de dados adequado.**

I. INTRODUÇÃO

A previsão de indicadores como o Índice de Desenvolvimento Humano Municipal (IDHM) é um desafio significativo na modelagem preditiva devido à complexidade das variáveis socioeconômicas envolvidas. Nos últimos anos, o aprendizado de máquina tem emergido como uma abordagem poderosa para lidar com problemas de regressão complexos, oferecendo soluções capazes de capturar padrões mais multifacetados e realizar previsões mais precisas. Entre as diversas técnicas disponíveis, modelos como Regressão Linear, Árvores de Decisão, Random Forest e Redes Neurais se destacam pela versatilidade e ótimo desempenho em problemas preditivos.

Este artigo tem como objetivo realizar uma análise comparativa entre diferentes algoritmos de regressão usados em aprendizado de máquina para prever valores contínuos em problemas reais. Técnicas menos sofisticadas, tais como a Regressão Linear e KNN, serão comparadas com técnicas mais sofisticadas, como as Árvores de Decisão e Random Forest; a fim de comparar suas forças e fraquezas em termos de precisão e acurácia, além de análisar sua robustez, capacidade de generalização e seu custo benefício.

A análise visa contribuir para a escolha de algoritmos mais adequados em tarefas de regressão, fornecendo insights sobre quais técnicas de aprendizado de máquina são mais eficazes e seus custos para a resolução de problemas preditivos complexos.

II. DESENVOLVIMENTO DOS MODELOS

A. Dataset escolhido

A escolha do conjunto de dados é um aspecto crucial para o sucesso de qualquer análise usando algoritmos preditivos. Portanto, o dataset utilizado, composto por dados socioeconômicos de cidades brasileiras para prever o Índice de Desenvolvimento Humano Municipal (IDHM), foi selecionado devido à sua alta complexidade e relevância. O IDHM é uma métrica que engloba diferentes dimensões de desenvolvimento, como educação, longevidade e renda, cada uma influenciada por um conjunto diverso de fatores. Essa multidimensionalidade torna o problema de previsão especialmente desafiador, necessitando de um bom tratamento de dados prévio, além de exigir modelos capazes de capturar relações complexas e interações não lineares entre as variáveis.

As cidades brasileiras têm perfis socioeconômicos extremamente diversos, com disparidades acentuadas entre regiões mais desenvolvidas e áreas menos favorecidas. Essa heterogeneidade introduz desafios adicionais para os modelos de aprendizado de máquina, uma vez que eles precisam lidar com diferentes distribuições de dados e capturar padrões complexos que podem não ser evidentes em abordagens tradicionais.

A escolha de um dataset com esse nível de complexidade é fundamental para testar a robustez e a capacidade dos modelos de aprendizado de máquina. A capacidade de lidar com dados de natureza tão diversa e altamente correlacionada é um indicador do quão bem um modelo pode ser aplicado a problemas reais de previsão. Desta forma, a utilização desse conjunto de dados não apenas enriquece a análise comparativa dos modelos de regressão, mas também torna os resultados mais relevantes e aplicáveis em cenários reais.

B. Tratamento de Dados

O tratamento de dados é uma etapa essencial para o uso de modelos aprendizado de máquina, especialmente em problemas de regressão complexos, como a previsão do Índice de Desenvolvimento Humano Municipal (IDHM). Antes de aplicar qualquer modelo, os dados brutos geralmente precisam passar por uma série de transformações para garantir que estejam prontos para análise e para maximizar o desempenho dos algoritmos.

No caso do dataset socioeconômico utilizado, o tratamento de dados envolveu várias etapas:[1]

1) **Limpeza de Dados**:

[1]Nem todos os métodos foram usados em todos os modelos.

- Foi necessário identificar e lidar com dados faltantes (valores ausentes), que podem distorcer as previsões dos modelos. Para o caso do dataset escolhido, não havia a presença de linhas com dados faltantes.
- A detecção e correção de *outliers* também foi realizada, uma vez que valores extremamente discrepantes podem afetar negativamente os modelos de regressão, especialmente em métodos sensíveis a variáveis numéricas, como a Regressão Linear.

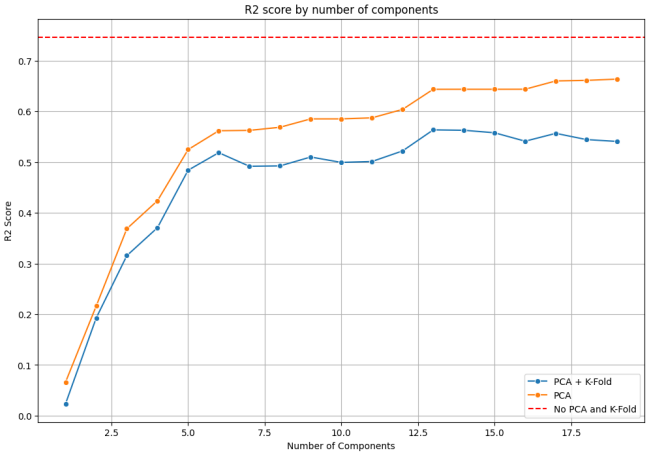
2) **Normalização e Padronização**: Como os algoritmos de aprendizado de máquina podem ser sensíveis às escalas das variáveis, especialmente os baseados em distância, como K-Nearest Neighbors (*KNN*), foi necessário normalizar as variáveis do dataset. A normalização ajusta os dados para que todos os valores fiquem entre um mesmo "range", permitindo que dados com números maiores ou menores não afetem o desempenho dos modelos.
3) **Codificação de Variáveis Categóricas**: Variáveis categóricas, como o estado (GO, MG etc), precisaram ser codificadas em um formato que os algoritmos possam processar. Foi usada a codificação ordinal, que atribui números às categorias.
4) **Redução de Dimensionalidade**: Para evitar o problema da "maldição da dimensionalidade", que pode afetar modelos como KNN e Random Forest, foram utilizadas técnicas de redução de dimensionalidade, como *Principal Component Analysis* (PCA) ou seleção de características baseada na importância das variáveis, para reduzir o número de atributos sem perder informação significativa aos modelos.
5) **Divisão do Dataset**: O dataset foi dividido em conjuntos de treino e teste, seguindo uma proporção típica de 80-20, para garantir que os modelos pudessem ser avaliados adequadamente quanto à sua capacidade de generalização. Ademais, foi utilizado o método de validação cruzada para garantir que os resultados não fossem enviesados por uma única divisão de dados.

O objetivo de todas essas técnicas de pré-processamento de dados foi garantir que os modelos pudessem trabalhar de maneira eficiente com os dados e capturar as complexas interações necessárias para prever o IDHM de maneira precisa. A qualidade do tratamento de dados tem um impacto significativo no desempenho dos algoritmos e pode fazer a diferença entre um modelo subótimo e um altamente eficaz.

C. Construção e Desempenho dos Modelos

1) Regressão Linear: O modelo de Regressão Linear teve sua testagem desenvolvida usando todos os métodos anteriormente citados para tratamento de dados. Foram executados testes nas seguintes condições:

- Alterando o número de componentes do PCA e alterando o número de k-folds para a validação cruzada;
- Alterando apenas o número de componentes do PCA, sem a utilização de validação cruzada;



- Alterando somente o número de k-folds para a validação cruzada, sem a utilização de PCA;
- Não utilizando de PCA e nem de validação cruzada.

O PCA tem seu número de componentes variando de 1 - 19 e o número de k-folds para a validação cruzada varia de 2 - 30.

É possível visualizar na Fig. II-C.1 que os melhores resultados foram os que não realizaram a redução de dimensionalidade e que não fizeram uso da validação cruzada; os resultados no modelo que utilizam apenas da validação cruzada são ainda piores, com melhor resultado de R^2 de 0.3629, com 5 folds.

O melhor resultado do modelo é o que não utiliza das técnicas citadas, tendo um resultado de R^2 de 0.7468.

2) K-Nearest Neighbors: O modelo K-Nearest Neighbors (*KNN*) sofreu de um processo de testagem similiar ao de Regressão Linear tendo a adição da variação da variável k, indicativa de com quantos vizinhos o modelo deve comparar a entrada:

- Alterando o número de componentes do PCA e alterando o número de k-folds para a validação cruzada;
- Alterando apenas o número de componentes do PCA, sem a utilização de validação cruzada;
- Alterando somente o número de k-folds para a validação cruzada, sem a utilização de PCA;
- Não utilizando de PCA e nem de validação cruzada.

O PCA teve seu número de componentes variando de 1 - 19, o número de k-folds para a validação cruzada varia de 2 - 30 e o número k de vizinhos passado ao modelo variou de 1 - 39.

Os resultados se demonstram como os piores, sendo o melhor modelo com 5 componentes para o PCA, sem a utilização de k-folds e com k número de vizinhos de 5, apresentando um R^2 de 0.3919.

3) Árvore de Decisão: O modelo que utiliza de Árvore de Decisão, especificamente a Árvore de Regressão para o nosso problema, não precisou realizar a normalização dos dados, devido ao poder da árvore de não ser influenciada por dados com escalas diferentes; e em alguns testes, a remoção de outliers também não foi utilizada.

Realizou-se os seguintes testes:

- Alterando o número de componentes do PCA e mantendo os dados outliers;
- Alterando o número de componentes do PCA e removendo os dados outliers;
- Não utilizando o PCA e mantendo os dados outliers;
- Não utilizando o PCA e removendo os outliers.

O PCA teve seu número de componentes variando de 1 - 19, o número de k-folds para a validação cruzada varia de 2 - 30, além disso, os seguintes hiperparâmetros foram alterados a fim de encontrar a melhor árvore possível:

- Profundidade máxima da árvore (5, 7 e 10);
- Número mínimo de amostras para dividir um nó (10, 20 e 50);
- Número mínimo de amostras em uma folha (1, 2 e 3);
- Critério de avaliação (erro quadrático médio de Friedman, erro absoluto e erro quadrático).

Os resultados encontrados foram mais interessantes que os modelos anteriores, mais robustos (*Linear Regression* e *KNN*). O melhor resultado encontrado foi o de uso de PCA com 4 componentes, com a remoção dos outliers, uso do critério de erro absoluto, profundidade máxima de 7 nós, número mínimo de amostras para dividir o nó de 2 e número minímo de amostras em uma folha de 10; tendo um resultado de R^2 de 0.7874, já bastante satisfatório dado o escopo da problemática.

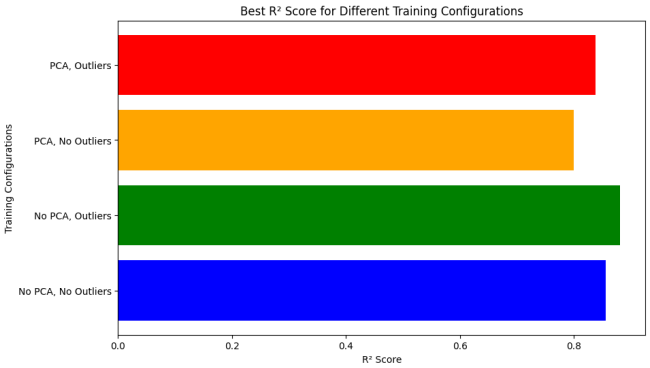
4) Floresta Aleatória: O modelo Floresta Aleatória (*Random Forest*) também, tal como a Árvore de Decisão, não necessitou de realizar a normalização dos dados e na testagem, uma abordagem mais ampla foi utilizada para encontrar resultados melhores.

Foi utilizado as seguintes variações de hiperparâmetros para a testagem do modelo (com tamanho fixo de k-fold de 5 para a validação cruzada):

- Número de árvores na floresta (45, 50, 65, 100, 150, 200, 250 e 300);
- Profundidade máxima da árvore (nenhum limite, 5, 10, 20 e 30);
- Número mínimo de amostras para se dividir um nó (2, 5, 10 e 20);
- Número mínimo de amostras em uma folha (1, 5, 10, 25 e 50);
- Número de variáveis (features) consideradas ao procurar a melhor divisão em cada nó (nenhum, raíz quadrada e $\log_2$).

Em suma, os resultados apresentados foram bastante abaixo do esperado para a complexidade do modelo, sendo o melhor resultado com 3 k-folds para a validação cruzada e com os seguintes hiperparâmetros: número de árvores na floresta de 300, profundidade máxima da árvore de 10, número mínimo de amostras para se dividir um nó de 2, número mínimo de amostras em uma folha de 1 e nenhuma variável foi utilizada para procurar a melhor divisão do nó; atingindo um valor de R^2 de 0.5430.

5) Árvores Extremamente Aleatórias: O modelo de Árvores Extramamente Aleatórias (*Extra Trees*) foi o que



acabou por obter o melhor resultado dentre os modelos apresentados. No modelo se realizou os seguintes testes

- Com o uso de PCA e com a remoção de dados outliers.
- Com o uso de PCA e sem a remoção de dados outliers
- Sem o uso de PCA e com a remoção de dados outliers
- Sem o uso de PCA e sem a remoção de dados outliers.

Todos os testes contaram com a normalização prévia dos dados. Além da variação dos seguintes hiperparâmetros:

- Número de árvores na floresta aleatória (300 e 500);
- Profundidade máxima da árvore (30 e 50);
- Número mínimo de amostras para se dividir um nó (10 e 20);

Como préviamente mencionado, foi o melhor modelo encontrado, tendo um valor de R^2 de 0.8816; sem a utilização de PCA e com a remoção dos outliers, e com os seguintes hiperparâmetros: número de árvores na floresta aleatória de 500, profundidade máxima da árvore de 50 e número mínimo de amostras para se dividir um nó de 10.

III. CONCLUSÃO

Finalmente, no decorrer deste artigo, foram avaliadas tanto técnicas mais simples, como Regressão Linear e KNN, quanto métodos mais avançados, como Árvores de Decisão, Florestas Aleatórias e Árvores Extremamente Aleatórias.

Vemos que a Regressão Linear, embora simples e amplamente utilizada, apresentou resultados bons, com um melhor R^2 de 0.7468 quando não foram aplicadas técnicas de validação cruzada ou redução de dimensionalidade. Entretando, o KNN mostrou-se menos eficaz, com um R^2 de apenas 0.3919, destacando sua sensibilidade às escolhas de hiperparâmetros e à normalização de dados, sendo a escolha do k vizinhos, um dos parâmetros mais importantes ao modelo.

Para mais, modelos mais complexos, como Árvores de Decisão e Florestas Aleatórias, trouxeram melhorias notáveis no desempenho. A Árvore de Decisão alcançou um R^2 de 0.7874, enquanto a Floresta Aleatória, mesmo sendo uma técnica robusta, apresentou resultados inferiores ao esperado, com um R^2 máximo de 0.5430.

O modelo que obteve o melhor desempenho foi o de Árvores Extremamente Aleatórias, que atingiu um R^2 de 0.8816, demonstrando ser a técnica mais adequada para a previsão do IDHM no contexto do dataset utilizado. Esse

resultado foi obtido sem o uso de PCA, mas com a remoção de outliers, reforçando a importância do tratamento prévio de dados para maximizar o potencial dos modelos de aprendizado de máquina.

Portanto, com base nos resultados apresentados, concluise que, em problemas de previsão de indicadores socioeconômicos complexos como o IDHM, técnicas baseadas em Árvores, nesse caso em especial, as Árvores Extremamente Aleatórias, são mais eficazes, especialmente quando combinadas com um tratamento de dados adequado. Esses modelos não apenas capturam melhor a complexidade das interações entre as variáveis, tem melhor performance com a grande dimensionalidade dos dados, mas também apresentam maior robustez em termos de generalização para novos dados.

REFERÊNCIAS

[1] “Kaggle.” https://www.kaggle.com/datasets/crisparada/brazilian-cities. Acessado em: 16 de Setembro de 2024.

[2] ”SciKit-Learn.” https://scikit-learn.org/stable. Acessado em: 22 de Setembro de 2024.

[3] ”BRAINS: Sua comunidade brasileira de Inteligência Artificial e Dados.” https://brains.dev/2023/extra-trees-arvores-extremamente-aleatorias. Acessado em: 23 de Setembro de 2024.